Reliable K-means clustering for data mining

Susanne Still
University of Hawaii at Mānoa
Information and Computer Sciences

Collaborators: William Bialek and Léon Bottou

K-means algorithm:

1. Initialize K cluster centers.
2. Assign each point to nearest cluster center.
3. Re-compute cluster centers as the means of the points assigned to each cluster.

The within-cluster sum of squares decreases in each iteration.

Problem: Local minima exist. No guarantee to find the global optimum. Sensitive to choice of initial conditions.

Soft $K$-means derived from rate–distortion theory:
Minimize the Lagrangian $F = I(c, i) - \beta \langle d(x_i, x_c) \rangle$ with respect to the assignment probabilities $p(c|i)$

$$\min_{p(c|i)} [I(c, i) - \beta \langle d(x_i, x_c) \rangle]$$

(1)

⇒ Optimal assignment rule (Gibbs distribution)

$$p(c|i) = \frac{1}{Z(i, \beta)} \exp \left[ -\beta d(x_i, x_c) \right]$$  (2)

with partition function

$$Z(i, \beta) = \sum_c \exp \left[ -\beta d(x_i, x_c) \right].$$  (3)

$\beta$: inverse temperature, if Lagrangian is interpreted as Free energy.

Minimize $F$ with respect to cluster centers, set gradients to zero ⇒

$$0 = \sum_i p(i|c) \frac{\partial d(x_i, x_c)}{\partial x_c}$$  (4)

For squared error distortion

$$x_c = \sum_i x_i \ p(i|c)$$  (5)
The **Information Bottleneck** (IB) method (N.Tishby, F.Pereira, W. Bialek, 1999) approaches clustering/lossy data compression, taking the view that a signal should be compressed in such a way that the compression contains maximal information about some variable of interest.

This contrasts with **rate distortion theory** (C. E. Shannon 1948), where the data is compressed such that the average distortion is minimized. Here, the resulting compression depends on the choice of the distortion function, while in the IB approach, the compression depends explicitly on the choice of a relevant variable.

The IB method is “distributional clustering”. Data with similar conditional distributions of the relevant variable given the data will be grouped together. Once this relevant variable is specified, there is no need to define an ad hoc measure of similarity, distance or distortion.
The IB approach has **conceptual advantages**, because the same information theoretic principle can be applied to the analysis of very different types of data, i.e. the analysis of both raw acoustic waveforms and of the words for which those sounds might stand (W. Bialek, 2001). Furthermore, this approach allows a principled treatment of questions which are often elusive in cluster analysis, for example how many clusters to use (S. Still and W. Bialek, 2004).

However, assuming that the data are given as points in a space with given distance measure, it is not clear how to incorporate a geometric intuition into the IB approach.
A different information theoretic treatment of clustering.

Intuitively, in geometrical problems, we want to cluster the points such as to keep maximal information about their location, i.e. location is the variable of interest. However, such a problem has a massively degenerate set of solutions, because information is a geometric invariant.

We propose to lift this degeneracy by choosing initial conditions which express our notion of geometry on the data space for the iterative algorithm that solves the IB.

This results in a new algorithm, which converges to the global optimum reliably and in reasonable time (faster than annealing).
Given: Data points $x_i$, $i = 1, \ldots, N$. Distance measure $d(x_i, x_j)$.

For simplicity, discretize space. $p(x|i) = \delta_{xx_i}$.

**Information Bottleneck approach:** Compress the data indices $i \in [1, N]$ into a smaller set of cluster indices $c \in [1, K]$ preserving as much information as possible about the location $x$ of the points,

$$\max_{p(c|i)} [I(x, c) - \lambda I(c, i)]$$

$\lambda$: Lagrange parameter. Regulates the trade-off between compression and preservation of relevant information.
Optimal assignment rule given by self-consistent equations which can be solved iteratively:

\[
p_n(c|i) = \frac{p_n(c)}{Z_n(i, \lambda)} \exp \left[ \frac{1}{\lambda} \sum_x p(x|i) \log_2 (p_n(x|c)) \right] \tag{7}
\]

\[
p_{n+1}(c) = \frac{1}{N} \sum_i p_n(c|i) \tag{8}
\]

\[
p_{n+1}(x|c) = \frac{1}{Np_n(c)} \sum_i p(x|i)p_n(c|i) \tag{9}
\]

\[
Z_n(i, \lambda) = \sum_c p_n(c) \exp \left[ \frac{1}{\lambda} \sum_x p(x|i) \log_2 (p_n(x|c)) \right] \tag{10}
\]
To lift the degeneracy, **initialize**

\[
p_0(x|c) = \frac{1}{Z_0(c)} \exp \left[ -\frac{1}{s} d(x, x_c^{(0)}) \right]
\]  

(11)

$Z_0(c)$ normalization constant; $s > 0$ arbitrary length scale.

In each iteration $n \geq 1$, determine the **cluster centers** $x_c^{(n)}$ according to

\[
0 = \sum_x p_n(x|c) \left. \frac{\partial d(x, x_c^{(n)})}{\partial x_c^{(n)}} \right|_{x_c^{(n)}}
\]  

(12)

For squared distance this reduces to

\[
x_c^{(n)} = \sum_x x \ p_n(x|c).
\]  

(13)

Furthermore initialize $p_0(c) = 1/K$. 

For a large range of temperatures $\lambda$, this converges to the "hard" $K$–means solution:

**Proposition:** For $\lambda < 1$ and $n \rightarrow \infty$

$$p(c|i) = \delta_{cc_i^*}. \quad (14)$$

with

$$c_i^* := \arg\min_c d(x_i, x_c). \quad (15)$$

where $x_c := x_c^\infty$.

For squared distance,

$$x_c = \frac{1}{n_c} \sum_i x_i \delta_{cc_i^*} \quad (16)$$

(This follows from substituting (14) into (13), using (8), (9) and $n_c = \sum_i \delta_{cc_i^*}$.)
Sketch of proof of Eqs. (14) and (15):

Substitute $p_0(x\mid c)$ into (7) ⇒ $p_0(c\mid i) \propto \exp\left[-\frac{1}{s\lambda}d(x_i, x_c^{(0)})\right]$. Substitute this back into (9) ⇒ $p_1(x\mid c) \propto \exp\left[-\frac{1}{s\lambda}d(x, x_c^{(0)})\right]$. Continue iteratively ⇒ accumulate factor $1/\lambda$ in the exponent in each iteration. After $n$ iterations:

$$p_n(c\mid i) = \frac{1}{Z_n(i, \lambda)} \exp\left[-\frac{1}{s\lambda(n+1)}d(x_i, x_c^{(n)})\right], \quad (17)$$

$$p_{n+1}(x\mid c) = \frac{1}{Z_n(c, \lambda)} \exp\left[-\frac{1}{s\lambda(n+1)}d(x, x_c^{(n)})\right]. \quad (18)$$

Therefore:

$$p_n(c_i^*\mid i) = \left(1 + \sum_{c \neq c_i^*} \exp\left[-\frac{1}{s\lambda(n+1)}\left(d(x, x_c^{(n)}) - d(x, x_{c_i^*}^{(n)})\right)\right]\right)^{-1} \quad (19)$$

$\forall c \neq c_i^*$: By definition $d(x, x_c^{(n)}) - d(x, x_{c_i^*}^{(n)}) > 0$, and thus

$$\exp\left[-\frac{1}{s\lambda(n+1)}\left[d(x, x_c^{(n)}) - d(x, x_{c_i^*}^{(n)})\right]\right] \xrightarrow{n \to \infty} 0 \quad (20)$$

Hence, the sum in eq. (19) tends to zero and we have $p(c_i^*\mid i) = 1$. Normalization ⇒ $p(c \neq c_i^*\mid i) = 0$. 
**For $\lambda = 1$, we obtain “soft $K$–means”**

Fixed point solution

$$p_n(c|i) = p(c|i) = \frac{1}{Z(i)} \exp \left[ -\frac{1}{s} d(x_i, x_c) \right] \quad \forall n, \quad (21)$$

with normalization $Z(i) = \sum_c \exp \left[ -\frac{1}{s} d(x_i, x_c) \right]$.

For squared distance,

$$x_c = \sum_i x_i \, p(i|c) \quad (22)$$

These are precisely the equations for the deterministic annealing approach to geometric clustering or “soft $K$–means” (K. Rose, E. Gurewitz and G. C. Fox, 1990).

**Interpretation:** The length scale $s$ appears in the position of the annealing temperature! This means that lowering the annealing temperature, in our view, is equivalent to decreasing the scale at which differences are taken seriously, in other words the smoothness scale.
That means that, in the regime $\lambda < 0$, $\lambda$ is the rate at which the “temperature” is lowered.

We derived an algorithm where the temperature is lowered \textbf{in each iteration}. Compare that to adiabatic changes in annealing $\Rightarrow$ we save time.

$\lambda$ is the parameter that controls the trade-off between compression and keeping relevant information. Hence $\lambda$ should be small to keep maximum relevant information! But that also makes the algorithm converge faster!

$s$ is the initial temperature and should be as large as possible to implement a non-informative prior.

$s$ is also the variance of the initial (gaussian) cluster distribution $\Rightarrow$ very large $s$ maximizes the uncertainty over the initial centers. This makes the algorithm stable to the actual choice of initial centers (initial conditions)!

Choose $\lambda \rightarrow 0$ and $s \rightarrow \infty$ $\Rightarrow$ \textbf{no parameter tuning}!

Get an algorithm that is both fast and reliable.
Example:

2500 data points drawn i.i.d from four Gaussian distributions with different means and the same variance. Result of numerical iteration of the equations (13), (17) and (18), with $\lambda = 0.5$ and $s = 0.5$. The algorithm converges to a stable solution after $n = 14$ iterations.
Robustness of algorithm to initial center positions as a function of the initial variance, $s$. 1000 different random initial positions were used to obtain clustering solutions on the data shown in previous figure.

Displayed is, as a function of the initial variance $s$, the percent of initial center positions that converge to a global maximum of the objective function.

In comparison, regular K–means (J. MacQueen 1967) converges to the global optimum for only 75.8% of the initial center positions.

The parameter $\lambda$ is kept fixed at reasonably small values (indicated in the plot) to ensure reasonably fast convergence (between 10 and 20 iterations).
Conclusion  A novel information theoretic approach to clustering in metric spaces (when the metric is given) allows one to

- derive both "soft" and "hard" K-means, depending on the value of the parameter that controls the trade-off between compression and the preservation of relevant information.

- "soft" K-means appears for a critical value of this parameter.

- In a large regime, the solution converges to the "hard" K-means solution.

- Can re-interpret the annealing temperature.

- The above derivation suggests a new algorithm with a fast cooling scheme.

- Good robustness at reasonable speed (faster than annealing).
References


